# Generative Entity-to-Entity Stance Detection with Knowledge Graph Augmentation

**Xinliang Frederick Zhang**[1], **Nick Beauchamp**[2], and **Lu Wang**[1]

[1]Computer Science and Engineering, University of Michigan, Ann Arbor, MI
[2]Department of Political Science, Northeastern University, Boston, MA
[1]{xlfzhang,wangluxy}@umich.edu
[2]n.beauchamp@northeastern.edu

## Abstract

Stance detection is typically framed as predicting the sentiment in a given text towards a *target entity*. However, this setup overlooks the importance of the *source entity*, i.e., who is expressing the opinion. In this paper, we emphasize the need for studying interactions among entities when inferring stances. We first introduce a new task, **entity-to-entity (E2E) stance detection**, which primes models to identify entities in their canonical names and discern stances jointly. To support this study, we curate a new dataset with 10,619 annotations labeled at the *sentence-level* from news articles of different ideological leanings. We present a novel generative framework to allow the generation of canonical names for entities as well as stances among them. We further enhance the model with a graph encoder to summarize entity activities and external knowledge surrounding the entities. Experiments show that our model outperforms strong comparisons by large margins. Further analyses demonstrate the usefulness of E2E stance detection for understanding media quotation and stance landscape, as well as inferring entity ideology.

## 1 Introduction

News media often employ ideological language to sway their readers, including criticizing entities they disagree with, and praising those conforming to their values (Baum and Groeling, 2008; Levendusky, 2013). However, in many cases, the sources do not directly express their sentiments, in part due to the norm that "objective" news media should restrict their role to narrating events and quoting others. In the realm of political news, many reported events consist of individuals or groups who themselves are engaged in praise or blame. The seemingly neutral act of choosing *who to quote*, and *about what*, as illustrated in Fig. 1, may be shaped by ideology and have significant effects on readers (Gentzkow and Shapiro, 2006; Gentzkow et al., 2015).

Trump's rhetoric, including calling Central Americans trying to enter the United States "an invasion," and his hard-line immigration policies have exposed him to condemnation since the El Paso shooting. "How far is it from Trump's saying this 'is an invasion' to the shooter in El Paso declaring 'his attack is a response to the Hispanic invasion of Texas?' Not far at all," Biden was due to say, according to an advance copy of his speech. "In both clear language and in code, this president has fanned the flames of white supremacy in this nation."

```
[0] Joe Biden  NEG  Donald Trump
[1] Joe Biden  NEG  white supremacy
[2] Donald Trump  POS  white supremacy
```

Figure 1: Sample stance triplet annotations for a target sentence. Entities in SEESAW can be Person or Topic, and are annotated in canonical forms. Multiple stances are expressed, whose inference needs context information, e.g., "president" refers to Donald Trump.

There thus exists a pressing need to examine these expressions of support and opposition in news articles (West et al., 2014) in order to understand how even apparently nonpartisan media can bias readers via the selective inclusion of stances among entities. Recognizing stances among political entities and events is also important in its own right: if copartisans are more likely to be positive towards each other and vice versa for counter-partisans, this allows us to (1) propagate partisanship and ideology through the signed network (De Nooy and Kleinnijenhuis, 2013), (2) infer ideology not just for politicians, but for events or objects (e.g., a new bill) that may inherently support the positions of specific groups (Diermeier et al., 2012), and (3) illuminate the implicit ideology of a journalist or media outlet (Hawkins and Nosek, 2012).

As a first step towards these goals, this paper presents the first study on solving the task of **entity-to-entity (E2E) stance detection** in an end-to-end fashion: Given a target sentence and its surrounding context, we extract a sequence of stance triplets that can be inferred from the input. Each triplet consists of a *source entity* and their *sentiment* towards a *target entity*, with entities in their canonical forms.

Existing stance detection methods are largely designed to infer an author's overall sentiment towards a given entity (Sobhani et al., 2017; Li et al., 2021a) or topic (Vamvas and Sennrich, 2020; Allaway and McKeown, 2020) from a text. E2E stance detection, by contrast, presents a number of new challenges. First, entities can be involved in multiple and even conflicting sentiments within a sentence,[1] as demonstrated in Fig. 1, suggesting the need to develop a model that can disentangle entity interactions. Second, entities are mentioned in various forms, e.g., full names or pronouns. Simply extracting the mentions would cause ambiguity for downstream applications. Canonical names that can be identified via knowledge bases (Shen et al., 2015) are thus preferred, which further requires the model to consider contextual information and global knowledge.

In this work, we first collect and annotate an E2E stance dataset, **SEESAW**[2] (Stance between Entity and Entity Supplemented with Article-level vieWpoint), based on 609 news articles crawled from AllSides.[3] SEESAW contains 10,619 stance triplets annotated at the sentence level, drawn from 203 political news stories, with each "story" consisting of 3 articles by media of different ideological leanings, as collected, coded, and aligned by AllSides. Our entities cover people, organizations, events, topics, and other objects.

We then present a novel encoder-decoder generative framework to output stance triplets in order. We first enhance the text encoder with a graph model (Velickovic et al., 2018) encoding a semantic graph that summarizes global entity interactions in the context, using relations extracted by an open information extraction (OpenIE) system (Stanovsky et al., 2018). On the decoder side, we improve the transformer decoder block (Vaswani et al., 2017) with a task-customized joint attention mechanism to combine textual and graph representations. Finally, external knowledge, such as party affiliation or employment, is injected into the graph encoder by adding extra nodes initialized with pretrained representations from Wikipedia. This allows the system to better characterize entity relations.

We conduct experiments on the newly collected SEESAW to evaluate models' capability of generating stance triplets, and additionally evaluate on a task of stance-only prediction when pairs of entities are given. Our model outperforms competitive baselines on E2E stance detection by at least 21% (relatively, accuracy of 11.32 vs. 13.74), demonstrating the effectiveness of adding knowledge from context and Wikipedia via graph encoding. Our best model also outperforms its pipeline counterpart which first extracts entities and then detects sentiment. This highlights the end-to-end prediction capability of generative models. Finally, we demonstrate the usefulness of E2E stances for media stance characterization and entity-level ideology prediction. Notably, we find that 1) both left- and right-leaning media tend to quote more from the Republican politicians; and 2) there appears a *symmetrical asymmetry* in expressed stances: the left is balanced while the right is biased in terms of expressed positivity, but the reverse holds for negativity.

## 2 Related Work

### 2.1 Stance Detection

Two major types of stance detection are widely studied (Aldayel and Magdy, 2021): (1) *sentiment-based*, the goal of which is to uncover the implicit sentiment (favor/against) evinced in texts towards a target, which can be a person or a topic (Mohammad et al., 2016; Sobhani et al., 2017; Allaway and McKeown, 2020; Li et al., 2021a); (2) *position-based*, which concerns whether a text snippet agrees/disagrees with a given claim or a headline (Ferreira and Vlachos, 2016; Pomerleau and Rao, 2017; Chen et al., 2019; Hanselowski et al., 2019; Conforti et al., 2020). In this work, we focus on the **sentiment-based** stance detection. Existing datasets for stance annotations are mainly based on social media posts (Mohammad et al., 2016; Li et al., 2021a), making the assumption that the sentiment is always held by the author, thus ignoring source entity annotation. Overall, there are at least three limitations for the existing stance detection data: 1) Data samples are collected within a narrow time period, e.g., a year, about specific events (Sobhani et al., 2017; Li et al., 2021a). 2) Entities are annotated by their mentions only (Deng and Wiebe, 2015; Park et al., 2021), limiting their usage for downstream applications. 3) Annotations are only available at either sentence-level or article-level, but not both. By contrast, our data spans a

[1] This may be a common phenomenon due to the journalistic norm of providing "balanced" views (Marshall, 2005; Moss, 2017).

[2] Our data and code can be accessed at https://github.com/launchnlp/SEESAW.

[3] https://www.allsides.com

10-year range at both sentence- and article-levels, with entities coded using canonical names.

Methodologically, existing models are designed for detecting stances towards a specific target entity (Mohammad et al., 2016; Augenstein et al., 2016). However, early methods assume the target entities in test time have been seen during training (Du et al., 2017; Zhou et al., 2017; Xue and Li, 2018). More recent work uses Large Language Models (LLMs) to enable stance prediction on unseen entities (Li et al., 2021b; Glandt et al., 2021). The most similar work to ours are Zhang et al. (2020) and Liu et al. (2021), both using graph convolutional networks (Kipf and Welling, 2017) to add external knowledge. Our model is different in at least two key respects: (1) They use existing knowledge bases, e.g., ConceptNet (Speer et al., 2017), with limited coverage of political knowledge. We instead resort to entity interactions extracted from news articles. (2) All prior models are based on Transformer encoder (Vaswani et al., 2017) only, while we explore the generative power of encoder-decoder models to address the more challenging E2E stance detection task. Moreover, none of these methods detects multiple stances from the same input, a gap that this work aims to fill.

### 2.2 Generative Models for Classification Task

Applying generative models for classification has recently gained research attention, mainly enabled by the wide usage of generative models, especially the large pretrained language models (Brown et al., 2020; Raffel et al., 2020). The most significant advantage of using generative models for classification problems resides in the improved interpretability between label semantics and input samples (Yan et al., 2021; Zhang et al., 2021), especially for multi-label classification problems (Yang et al., 2018; Liao et al., 2020; Yue et al., 2021). Generative models are especially suitable for our task, since canonical names are preferred in the output. Recent work (Humeau et al., 2020; Cao et al., 2021) supports our assumption by showing that generative models are better at capturing fine-grained interactions between the text and entities than encoder-only models. This work carries out the first study of deploying such models for a new task, E2E stance detection. In addition, it extends the model with context information and external knowledge.

## 3 SEESAW Collection and Annotation

**Annotation Process.** We use AllSides news stories collected by Liu et al. (2022), where each story contains 1-3 articles on the same event but reported by media of different ideology. The stories span from 2012 to 2021. We only keep news stories with 3 articles and that are pertinent to U.S. politics. We manually inspect and select stories to cover diverse topics. The resulting SEESAW contains 52 topics (full list in Table A1). We further clean articles by removing boilerplate and noisy text.

We hired six college students with high English proficiency to conduct annotation tasks. Each person is asked to read all articles written on the same story before annotating. We summarize the main steps below, with detailed protocol in Appendix A.

1. They start with reading the article, and then identify entities of political interests that are involved in sentiment expressions. An entity can have a type of person, organization, place, event, topic, or religion. Annotated entities are categorized into *main* and *salient* entities.[4]

2. For *each sentence*, stance is annotated between entities in the triplet format, i.e., <`source`, `sentiment`, `target`> where sentiment can be either *positive* or *negative*. We use `Author` as the source entity, if no clear source entity is found. Also, `Someone` is used to replace source or target entities of no clear political interest, e.g., "a neighbor".

3. At the article level, we annotate its overall sentiment towards each identified entity, together with the ideology of each entity as well as the ideological leaning of the article, all in 7-way. We then convert the annotations on sentiment and ideological leaning into 3-way and 5-way, respectively.

Finally, we conduct cross-document entity resolution by linking annotations to their *canonical names* according to Wikipedia, e.g., "this president" becomes "Donald Trump" as in Fig. 1.

A quality control procedure is designed to allow annotators to receive timely feedback and improve agreement over time. Details are in Appendix C. Importantly, over 60 randomly sampled articles,

[4] *Main entities* are defined as main event enablers and participants as well as the ones that are affected by such events. *Salient entities* refer to other political or notable figures that appear in the news stories who are not the main entities.

the average overlap of annotated entities between pairs of coders is 55.5%. When both source and target entities match, the sentiment agreement level is 97%, indicating the high quality of the dataset.

**Statistics.** SEESAW includes 609 news articles from 203 stories, published by 24 different media outlets (9 left, 6 central, and 9 right). Table A2 lists all the media outlets. On average, each article contains 28 sentences and 647 words. 44% of sentences per article have at least one stance annotation, among which 29% have at least two.

In total, there are 10,619 **stance annotations** in SEESAW, covering 1,757 **distinct entities**. 62.4% of the triplets have *negative* sentiment. **Entity types** in SEESAW cover People (49.6%), Organization (12.8%), Place (4.2%), Event (12.0%), Topic (17.4%), Religion (0.2%), and Others (3.8%), showing its diversity of entity annotations. It is worth noting that the *source entity* being <*Author*> and <*Someone*> occurs 9.1% and 12.5% of the annotations. Meanwhile, the number for *target entity* being <*Someone*> is 5.3%.

In terms of **entity ideology**, the portions of entities annotated as liberal, moderate, and conservative is 31.0%, 15.9%, and 34.6%, respectively.[5] Our annotated **article leanings** align with AllSides' media-level labels for 76.7% of the time.

## 4 Generative E2E Stance Detection

Fig. 2 depicts the end-to-end generative framework that reads in a document and produces stance triplets in an auto-regressive fashion, by leveraging multiple knowledge sources using graph augmentation. We use BART (Lewis et al., 2020), a large pretrained encode-decoder model, as the backbone. Taking as input a *target sentence* from a document $\mathbf{d}$, our model first constructs a semantic graph $G$ (§4.1), which is encoded via the graph encoder (§4.2). The contextualized representations of tokens and nodes are denoted as $\mathbf{H}_T$ and $\mathbf{H}_G$. Stance triplets are generated by our decoder using improved *in-parallel attention* and *information fusion* mechanisms (§4.3). Moreover, we inject Wikipedia knowledge to support the identification of relations between entities (§4.5). Below we describe the main components, with additional formulation and implementation details in Appendix B.

[5]Some entities' ideologies are marked as N/A such as Kremlin (non-US entity) and amnesty (topic).

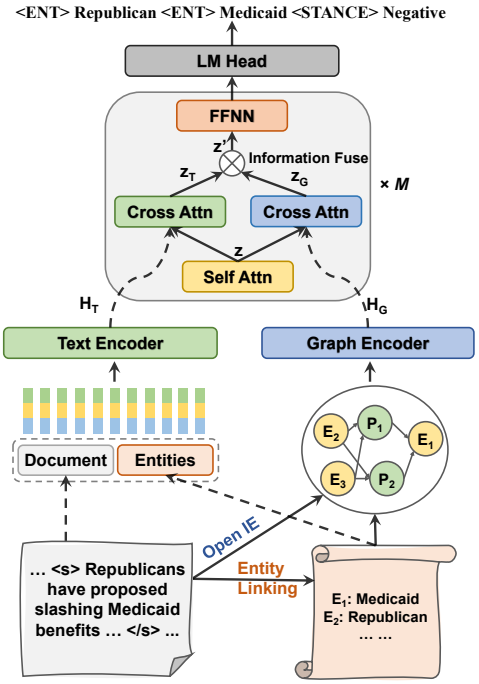


Figure 2: Overview of our end-to-end generative framework for stance detection. Our model reads a document $\mathbf{x}$, on which we construct a semantic graph $G$ (§4.1). $G$ contains three types of nodes: entity nodes $E_i$, predicate nodes $P_i$, and Wiki nodes (not shown in the diagram). Extracted entities are paired with document $\mathbf{x}$ and then fed into text encoder, in the format of "$\mathbf{x}$ <*s*> <*ENT*> $E_1$ <*ENT*> $E_2$ ...". Besides token and position embeddings, we add a third embedding to indicate a token's type: preceding context, target text, succeeding context, or entities. Our decoder implements in-parallel cross-attention (§4.3) to attend both text ($\mathbf{H}_T$) and node ($\mathbf{H}_G$) representations concurrently. Fused representations are obtained through the information fusion layer.

### 4.1 Local Semantic Graph Construction

Our goal is to construct a semantic graph that can summarize events and sentiments involving entities in the document context. We thus use OpenIE (Stanovsky et al., 2018) to obtain semantic relation outputs in the form of <subject, predicate, object>. Triplets whose span is longer than 15 tokens are dropped. We also conduct global entity linking[6] to extract canonical names for all entities in the document, which are fed into the text encoder as shown in Fig. 2. This linking step facilitates the generation of canonical names in a consistent manner.

We start constructing the graph $G$ by treating the extracted entities as *entity nodes*, where co-referential mentions of the same entity are collapsed into a single node. Following Beck et al. (2018) and Huang et al. (2020), we further create

[6]https://cloud.google.com/natural-language/

*predicate nodes*. We then add directed edges from subject to predicate and from predicate to object. We add reverse edges and self-loops to enhance graph connectivity and improve information flow.

### 4.2 Graph Encoder

We initialize node representations $\mathbf{H}_G$ by using the average contextual token embeddings ($\mathbf{H}_T$) of all co-referential mentions. Similar to Yasunaga et al. (2021), we add a *global* node, initialized with mean pooling over all tokens in the target sentence. The global node is connected to entity nodes in $G$ to allow better communication of text knowledge.

Our graph encoder improves upon Graph Attention Networks (GAT; Velickovic et al., 2018) using Transformer layer networks and *Add & Norm* structures (Vaswani et al., 2017). Concretely, in each layer, we use the multi-head GAT massage passing rule to update node representations $\mathbf{H}_G$, and then pass them through a 2-layer MLP. Residual connections (He et al., 2016) and layer normalization (Ba et al., 2016) are employed to stabilize the hidden state dynamics. We use two layers with 8-head GAT to produce final node representations $\mathbf{H}_G$.

### 4.3 Decoder

We further improve the Transformer decoder to enable reasoning over both the text and the graph. The key difference between the vanilla Transformer decoder and ours is the *in-parallel cross-attention* layer which allows better integration of knowledge encoded in the two sources. Concretely, in-parallel cross attentions are implemented as follows:

$$\begin{aligned} \mathbf{z}_T &= \text{LayerNorm}(\mathbf{z} + \text{Attn}(\mathbf{z}, \mathbf{H}_T)) \\ \mathbf{z}_G &= \text{LayerNorm}(\mathbf{z} + \text{Attn}(\mathbf{z}, \mathbf{H}_G)) \end{aligned} \quad (1)$$

where $\mathbf{z}$ denotes the output from the self-attention layer and $\text{Attn}(\cdot,\cdot)$ is the same cross-attention mechanism as implemented in Vaswani et al. (2017).

Next, we introduce an *information fusion* module to enable the interaction between textual ($\mathbf{z}_T$) and graph ($\mathbf{z}_G$) hidden states, in order to obtain the fused representation, $\mathbf{z}'$. We implement two information fusion operations: (1) **addition**, i.e., $\mathbf{z}' = \mathbf{z}_T + \mathbf{z}_G$, and (2) **gating** mechanism between $\mathbf{z}_T$ and $\mathbf{z}_G$ as in Zhao et al. (2018) except that we use $\text{GELU}(\cdot)$ as the activation function. The operation selection is determined by downstream tasks.

### 4.4 Training Objectives

We adopt the cross entropy training objective for model training, $\mathcal{L}_{stance}$. The reference $\mathbf{y}$ is a sequence of ground-truth stance triplet(s), sorted by their entities' first occurrences in the target sentence. Input and output formats are shown in Fig. 2.

**Variant with Node Prediction.** In addition to modeling entity interactions in the graph, we enhance the model by adding an auxiliary objective to predict node salience, i.e., whether the corresponding entity should appear in the stance triplets $\mathbf{y}$ to be generated. This is motivated by the observation that $G$ usually contains excessive entity nodes, only a small number of which are involved in sentiment expression in the target sentence. Specifically, for each entity node, we predict its salience by applying affine transformation over its representation $\mathbf{h}_G$, followed by a sigmoid function to get a single value in $[0, 1]$. We adopt the binary cross entropy (BCE) objective to minimize the loss $\mathcal{L}_{node}$ over all *entity nodes*. Finally, when the node prediction module is enabled, the overall loss for the **multitask** learning setup is $\mathcal{L}_{multi} = \mathcal{L}_{stance} + \mathcal{L}_{node}$.

### 4.5 Graph Expansion: Wiki Knowledge Injection

To gain a better understanding of stances among entities over controversial issues, it is useful to access external knowledge about the entities mentioned in the news, e.g., their party affiliations. Therefore, we obtain the knowledge representations for entities using Wikipedia2Vec (Yamada et al., 2020), a tool that jointly learns word and entity embeddings from Wikipedia. The learned entity embeddings are shown to be effective in encoding the background knowledge discussed in Wikipedia. These embeddings are then added as *wiki nodes* in graph $G$. We add edges between an entity node and a wiki node, if the entity is linked to the corresponding Wikipedia entry based on entity linking.[7] In our implementation, we take the pre-trained 500-dimensional vectors,[8] transformed by a two-layer MLP, for node representations initialization.

To summarize, graph-augmented generative models have been studied for several generation tasks, including abstractive summarization (Huang et al., 2020), question answering (Yasunaga et al., 2021), and question generation (Su et al., 2020). Prior design of graph encoders uses either external knowledge bases (Zhang et al., 2022) or a local graph constructed using semantic parsing (Cao and

[7] https://cloud.google.com/natural-language/

[8] wikipedia2vec.s3.amazonaws.com/models/en/2018-04-20/enwiki_20180420_500d.pkl.bz2

| | SEESAW (Task A) | Park et al. (Task B) |
|---|---|---|
| Target Sentence length | 30.3 | 31.0 |
| Label ratio (pos/neg) | 37.6%/62.4% | 35.3%/64.7% |
| Splits (train/valid/test) | 4505/1313/1378 | 4252/506/562 |

Table 1: Statistics of the two datasets for experiments. Data by Park et al. (2021) only contains single sentences without context. We split the SEESAW chronologically, and use the same splits as in Park et al. (2021).

Wang, 2021). Since large-scale structured knowledge base does not exist for the political domain, our method differs from previous work in that we leverage both entity interactions from the context and external knowledge from Wikipedia in a unified graph representation learning framework to better characterize entity interactions.

## 5 Experiments

### 5.1 Tasks and Datasets

We conduct evaluation on two different stance detection tasks. Table 1 shows the basic statistics of datasets and splits.

**Task A: Entity-to-entity stance detection.** We experiment with SEESAW for generating stance triplets of <source, sentiment, target>. The input text can be a target sentence alone or with surrounding context (up to $k$ preceding and $k$ succeeding sentences). We set $k = 3$ for all experiments.

**Task B: Stance-only prediction for pairwise entities.** Park et al. (2021) build a dataset annotating sentiments between mentions rather than canonical entities. We include this dataset to assess our model's capability on a stance-related classification task. For experiments, we only keep samples with positive and negative sentiments. Formally, their input contains one sentence $s$ and two entities $e_1$ and $e_2$. The goal is to jointly predict the direction and the sentiment, i.e., four labels in total.

### 5.2 Baselines

For Task A, since there is no existing E2E stance detection models, we consider finetuning BART using different inputs as baselines: (1) **sentence**: target sentence only; (2) **sentence + context**: target sentence with surrounding context; (3) **sentence + context + entities**: additionally appending all entities in their canonical names as extracted in §4.1, same as our model's input.

We further consider two variants of our model as baselines. We first design a **pipeline model**, which first uses the node prediction module to identify salient entities for inclusion in the stance triplets. Then we introduce a soft mask layer over entity nodes in $G$ before passing them into the graph encoder, by multiplying node representations with their predicted salience scores. We also experiment with **oracle entities**, where we feed in the ground-truth salient entities for text encoding and graph representation learning.

Finally, to test the effectiveness of our designed in-parallel cross-attention, we compare with a **sequential attention**, designed by Cao and Wang (2021), to consolidate text and graph modalities. They allow the decoder hidden states to first attend token and then node representations. Their model differs from our model only in the attention design.

On Task B, since LLMs have obtained the state-of-the-art performance on existing stance prediction tasks (Glandt et al., 2021; Liu et al., 2022), we compare with the following LLM-based methods in addition to **BART**. We compare with **DSE2QA** (Park et al., 2021), which is built on top of RoBERTa (Liu et al., 2019). They transform the sentiment classification task into yes/no question answering, where the questions ask whether a sentiment can be entailed from several manually designed questions appended after the target sentence. We use the question that obtained the best performance on their dataset, i.e., "$e_1$ - $e_2$ - [sentiment]". We then consider a recent LLM, **POLITICS** (Liu et al., 2022), trained on RoBERTa with ideology-driven pretraining objectives that compare articles on the same story.

### 5.3 Evaluation Metrics

For both tasks, we report accuracy and F1 scores. For Task A, accuracy is measured at the sample level, i.e., all stance triplets need to be generated correctly to be considered as correct. F1 is instead measured at the triplet level. We include another metric, accuracy-**any**, where for each sample, the prediction is considered correct if at least one triplet is found in the reference. We also break down the triplets, and evaluate varying **aspects** based on pairs of source entity-sentiment (**src-s**), sentiment-target entity (**s-tgt**), and source-target entities (**src-tgt**), using accuracy-any.

### 5.4 Results

Results for E2E stance detection is displayed in Table 2. Compared with baselines, we see significant improvements by providing context and entities in canonical forms, indicating that adding story context and additional knowledge about entity names

| | Full | | | Aspect | | |
|---|---|---|---|---|---|---|
| | Acc. | F1 | Acc.Any | src-s | s-tgt | src-tgt |
| **Baselines (no graph)** | | | | | | |
| Sentence (Sen) | 7.26 | 10.35 | 12.39 | 36.66 | 23.81 | 14.88 |
| Sen + Context (Ctx) | 9.66 | 14.08 | 16.87 | 45.24 | 27.79 | 20.01 |
| Sen + Ctx + Entities | 11.32 | 16.00 | 19.15 | 47.43 | 30.03 | 23.18 |
| **Pipeline Models (Ours)** | | | | | | |
| Graph | 12.03 | 15.77 | 19.86 | 46.60 | 31.07 | 23.19 |
| + Oracle Entities | 31.84 | 35.58 | 44.87 | 66.33 | 55.50 | 53.82 |
| **End-to-end Models (Ours)** | | | | | | |
| Graph (seq. attn.) | 12.97 | 17.22 | 20.62 | 50.58 | 32.45 | 24.76 |
| Graph | 13.62 | 18.12 | 21.78 | 51.01 | 32.65 | **26.08** |
| + Multitask | 13.34 | 18.16 | 21.77 | **52.10** | 32.06 | 26.07 |
| + Wiki | **13.74** | **18.24** | **21.87** | 51.41 | **32.69** | 25.94 |

Table 2: Results on SEESAW for E2E stance detection task, and breakdown of accuracy scores by aspects (average of 5 runs). Best results without oracle entities are in **bold**. Our graph-augmented model with Wikipedia knowledge performs the best on 4 out of 6 metrics, indicating the effectiveness of encoding knowledge. Results with standard deviation are in Table A3.

is useful for the E2E stance detection task.

Next, though the pipeline variant of our model provides better explainability as it first identifies salient entities, it yields inferior performance than the end-to-end version of our model. After inspection, we find that the salient entity prediction module only reaches around $58\%$ for F1. With the oracle entities as input, we see a significant boost in the performance, highlighting the importance and difficulty of entity understanding and extraction.

Importantly, our model enhanced with Wikipedia knowledge performs the best on 4 out of 6 metrics. This signifies the effective design of graph modeling on entity interactions. Moreover, our newly designed in-parallel attention is also shown to be more effective than attending the two sources of information in sequence, as done in Cao and Wang (2021). This implies that having symmetric integration of text and graph can be important, though this should be tested on other tasks in future work. When breaking down the predicted stance triplets into different pairs, we see that identifying source entity and sentiment is easier than predicting sentiment and target entity. This might be because target entities are often introduced earlier in the article, thus requiring long-term discourse understanding.

Finally, the overall performance of E2E stance detection is quite low for all models. This is mainly because models may fail to generate exactly the same canonical names for entities and often fall short of producing all stance instances when multiple sentiments are embedded in a single sentence.

| | Accuracy | Macro F1 |
|---|---|---|
| BART (Lewis et al., 2020) | 86.32 | 77.53 |
| POLITICS (Liu et al., 2022) | 86.33 | 77.48 |
| DSE2QA (Park et al., 2021) | 87.78 | **79.90** |
| Our Model | **87.79** | 79.01 |

Table 3: Results on stance-only prediction for specified pairwise entities. Our model performs on par with state-of-the-art models in stance detection tasks (POLITICS and DSE2QA). Results with std. deviation in Table A4.

On the stance-only prediction task, Table 3 shows that our model yields better or comparable performance than the state-of-the-art models. This demonstrates that our generative stance detection model can also perform well on a quaternary classification setup. Note that the input text is short ($\sim$30 tokens), limiting our model's capability of capturing global context. However, our model still outperforms BART and the recently trained LLM, POLITICS, designed for ideology prediction and stance detection. The experimental results are inline with Lewis et al. (2020) that the improvements on generation tasks do not come at the expense of classification performance.

### 5.5 Sample Output and Error Analysis

Fig. 3 shows one test example with system outputs from baselines and our models. All three baseline models detect a positive sentiment ascribed to Mike Pence but fail to uncover the specific target entity. However, our graph-augmented model manages to produce stance triplet [1], using the direct edge linking Kamala Harris and Donald Trump through a negative predicate in the corresponding graph $G$. We also find that our model using Wikipedia knowledge often uncovers hidden relations between entities, e.g., party affiliations and geopolitical relations, which are useful for stance detection but cannot be inferred from the document alone. Such as in this example, to enable the generation of stance triplet [0], our model leverages Wikipedia knowledge to draw the connection between Wisconsin and "*the state*" in the news. Moreover, this example showcases the power of our model in generating multiple stances, which is an essential capability for the E2E stance detection task. In Fig. A1, we show another example, where none of the models produces a correct stance triplet, further confirming the challenge posed by E2E stance detection. This points out the future direction of investigating more powerful models that can better make inferences and perform reasoning over knowledge.

Ms. Harris, who is making her first trip to a battleground state since joining the Democratic ticket, is visiting with union workers and leaders as well as African-American businesspeople and pastors in Milwaukee, the Black hub of the state. Each is expected to focus on the economy, with Mr. Pence hailing **the state**'s job growth before the coronavirus pandemic and Ms. Harris critiquing the administration's handling of the virus and the resultant impact on the economy. Yet their political missions are different. The vice president is hoping to appeal to voters in a historically Democratic part of **Wisconsin**, where Mr. Trump outperformed his Republican predecessors, in hopes they abandon their political roots again.

`[0] Mike Pence POS Wisconsin`
`[1] Kamala Harris NEG Donald Trump`
`[2] Kamala Harris NEG <Someone>`

---

**Sent.**: [0] Mike Pence POS <Someone>
**Sent. + Cxt.**: [0] Mike Pence POS <Someone>
**Sent. + Cxt. + Ent.**: [0] Mike Pence POS <Someone>
**Graph model** (ours): [0] Mike Pence POS job growth; [1] Kamala Harris NEG Donald Trump
**Graph model + Wiki** (ours): [0] Mike Pence POS Wisconsin; [1] Kamala Harris NEG Donald Trump

Figure 3: Sample system predictions (below the dotted line) with human labeled triples (above the dotted line). Target sentence is underlined. All three baselines fail to identify the correct target entity. By contrast, our graph-augmented end-to-end model predicts the first triplet by leveraging the semantic relation as captured by graph $G$. After encoding Wikipedia knowledge, our model draws the connection between Wisconsin and "the state" in the text, thus generating a correct stance triplet [0]. Our models also produce multiple triplets.

## 6 Further Analysis

In SEESAW, we are able to identify the partisanship of 204 politicians (Democrat vs. Republican) based on Voteview.[9] This subset accounts for more than 60% of person mentions in the dataset. We further include *Democrat* and *Republican* as two separate entities, since they are also frequently mentioned in news. Analyses in this section are done based on this entity set (henceforth *analysis set*).

### 6.1 Landscape of Media Quotation and Stance

We start with examining the relation between media ideology and their stances. We first study *do media tend to quote people of the same or opposite ideologies?* To answer this question, we count the average number of political figures quoted as source entities in each article. Interestingly, media from both sides are more likely to quote Republican politicians, as seen in Fig. 4. This is consistent with recent study on U.S. TV media (Hong et al., 2021), where the authors show that Republicans receive more screen time as well as get longer TV interviews time than Democrats.[10] Additionally, left-leaning outlets use more quotes than their right counterparts, which also aligns with previous observations (Welch et al., 1998).

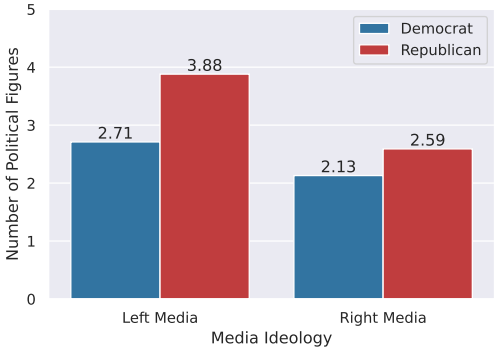


Figure 4: Media quoting Democrats vs. Republicans by counting source entities per article. Both left- and right-leaning media outlets quote Republicans more.

Next, we ask: *do media tend to be more positive towards people with the same ideology as theirs and be more negative towards out-group entities?* Here we consider stance triplets containing the target entities in the analysis set. First, as can be seen from Fig. 5, more negative sentiments are observed in news articles, which align with existing work that shows journalists more often report blame than praise (Damstra et al., 2020). More importantly, we observe an interesting sentiment pattern of **symmetrical asymmetry**: Left-leaning media produces articles use similar amounts of positivity towards Democrats and Republicans (15.4% vs. 15.8%), while right-leaning media are more positive towards Republicans (18.8% vs. 8.7%). By contrast, when it comes to negativity, right-leaning media are more balanced (36.4% vs. 36.1%), while left-leaning media are unbalanced (25.6% to Democrat vs. 43.2% to Republicans). This suggests that the left and right media may be biased in different ways: the left by directing more negativity to the opposing side, the right by directing more positivity towards their own side.

### 6.2 E2E Stances for Ideology Prediction

Here we test whether the knowledge of E2E stances can help with entity-level ideology prediction. Based on the sentiments expressed among politicians, we construct a directed graph with edges indicating the direction and sentiment between en-

[9] https://voteview.com/

[10] By original authors (Hong et al., 2021), the conclusion of interview time might not be true due to biased data sampling.

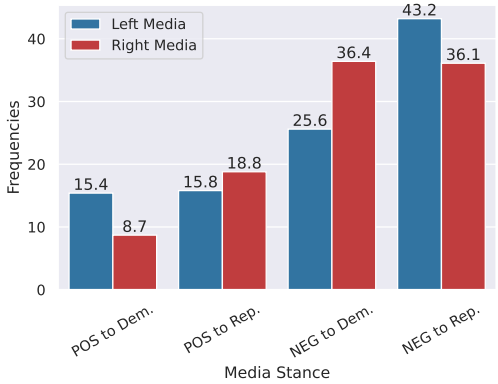


Figure 5: Percentage of stance triplets that media favoring or criticizing entities from the same or the opposite side. Media of both sides attack politicians from the opposite parties more than their own parties. Note there is a *symmetrical asymmetry* phenomenon: Left is balanced while the right is unbalanced in terms of indicated positivity, and the other way around for negativity.

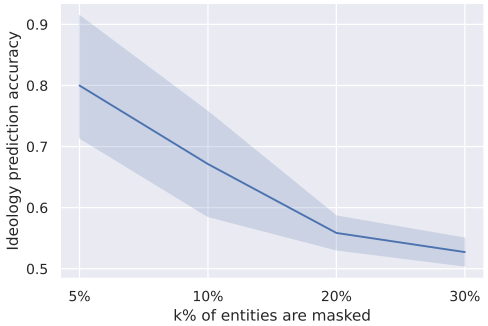


Figure 6: Entity-level ideology prediction using stances from/to their neighboring entities with known ideology. We increase the ratio of entities being masked, which decreases the ideology prediction accuracy. This implies knowing entity's support/oppose interactions with other entities is helpful for predicting their own ideology.

ties. We random mask the ideology of $k\%$ of the nodes, and then infer their ideology using sentiments from/to their neighbors. Each node $e$ has two counters: $c^D$ for Democrat and $c^R$ for Republican. For each of $e$'s neighbors with known ideology, $c^D$ increases by 1 if (1) the neighbor is D and positive sentiment is expressed in either direction, or (2) the neighbor is R and negative sentiment is observed in either direction. Similar rules are used for $c^R$. The counter with the higher value decides $e$'s ideology.

By varying the percentage of nodes to be masked (Fig. 6), we observe that, the more we know about an entity's sentiment interactions with others, the more accurate we can predict their ideology. This shows the usefulness of networks constructed from E2E stances as inferred from news articles.

### 6.3 Inter- and Intra-group Sentiment

Finally, we observe that the majority of inter-party stances are negative, e.g., $92.7\%$ of sentiment by Democratic politicians towards Republicans is negative, and the number of republicans is $91.9\%$. This is unsurprising given the current level of polarization in the U.S. (Campbell, 2018; Klein, 2020). Notably, the Republican Party is much more divided compared with Democrats, where more than half of intra-group stances (i.e., $56.0\%$) within Republican Party carry negative sentiments, whereas the percentage for Democrats is only $25.2\%$. These results contradict recent observations for in-group sentiment as measured on social media users and congressional members (Grossmann and Hopkins, 2016; Benkler et al., 2018). This highlights the significance of studying stances quoted by news media and suggests new avenues for future research.

## 7 Conclusion

We present and investigate a novel task: entity-to-entity (E2E) stance detection, with the goal of extracting a sequence of stance triplets from a target sentence by jointly identifying entities in their canonical names and discerning stances among them. To support this study, we annotate a new dataset, SEESAW, with 10,619 sentence-level annotations. We propose a novel end-to-end generative framework to output stance triplets. Specifically, we enhance standard encoder-decoder models with a semantic graph to capture entity interactions within context. We further augment our model with external knowledge learned from Wikipedia, yielding the best overall performance. We conduct further analyses to demonstrate the effectiveness of E2E stances on media landscape characterization and entity ideology prediction.

## Acknowledgments

This work is supported in part through National Science Foundation under grant IIS-2127747, Air Force Office of Scientific Research under grant FA9550-22-1-0099, and computational resources and services provided by Advanced Research Computing (ARC), a division of Information and Technology Services (ITS) at the University of Michigan, Ann Arbor. We appreciate the anonymous reviewers for their helpful comments. We thank David Wegsman, Isabella Allada, Margaret Peterson, Jiayi Zhang and Bingyan Hu for their efforts in SEESAW construction. We also thank Jiayi Zhang and Hongyi Yin for conducting data checking.

## Limitation

### GPU resources

The framework proposed in this work is an encoder-decoder based generative model. It is thus more time-consuming than standard discriminative models for training and evaluation, which in turn results in higher carbon footprint. Specifically, we run our experiments on 1 single NVIDIA RTX A6000 with significant CPU resources. The training time for our model is usually around 5 hours.

### System limitation

In spite of achieving the best performance on E2E stance detection and comparable performance with the SOTA model (Park et al., 2021) on the task of stance-only prediction given pairwise entities, our model is still limited in the following aspects. (1) From Table 2, even the best model struggles with extracting correct <source, target> pairs. (2) Though we have pre-processed the data and conducted global entity linking, which helps with entity-level coreference resolution, better designs are needed to help resolve event coreference. Concretely, as shown in Fig. A1, our best model still suffers from making sense of the correct relation between "such an action" and "affidavit".

### Evaluation limitation

We believe the high-quality annotations and diverse entities in our SEESAW can help foster research along this novel research direction. However, the adopted evaluation schemes still have their own shortfalls. For example, in Fig. 3, our model's output, *Mike Pence POS job growth*, can be considered as correct. Yet, under the current automatic evaluation scheme, this prediction is counted as a mistake. More robust and accurate metrics need to be developed to gauge the research progress.

## Ethical Consideration

### SEESAW

**SEESAW collection.** All news articles were collected in a manner consistent with the terms of use of the original sources as well as the intellectual property and the privacy rights of the original authors of the texts, i.e., source owners. During data collection, the authors honored privacy rights of content creators, thus did not collect any sensitive information that can reveal their identities. All participants involved in the process have completed human subjects research training at their affiliated institutions. We also consulted Section 107[11] of the U.S. Copyright Act and ensured that our collection action fell under the fair use category.

**SEESAW annotation.** In this study, manual work is involved. All the participants are college students, who participated in the this project for credits rather than compensation. We treat every annotator fairly by holding weekly meetings to give them timely feedbacks and grade them quite leniently to express our appreciation for their consistent efforts.

### Benefit and Potential Misuse of our developed Systems and SEESAW

**Intended use.** The models developed in this work can assist the general public to measure and understand stance evinced in texts. For example, our model can be deployed in wild environments to automatically extract stance triplets at no cost.

**Failure mode** is defined as situations where our model fails to correctly extract a stance triplet of a given text. In such cases, our model might deliver misinformation or cause misunderstanding towards a political figure or a policy. For vulnerable populations (e.g., people who maybe not be able to make the right judgements), the harm could be tremendously magnified when they fail to interpret the model outputs or blindly trust systems' outputs. Ideally, the interpretation of our model's predictions should be carried out within the broader context of the source text.

**Misuse potential.** Users may mistakenly take the machine prediction as a golden rule or a fact. We would recommend any politics-related machine learning models, including ours, put up an "use with caution" message to encourage users to check more sources or consult political science experts to reduce the risk of being misled by single source. Moreover, our developed system might be misused to label people with a specific stance towards an issue that they do not want to be associated with. We suggest that when in use the tools should be accompanied with descriptions about their limitations and imperfect performance, as well as allow users to opt out from being the subjects of measurement.

**Biases and bias mitigation.** No known bias is observed in SEESAW since we collected balanced views of news stories from AllSides. During annotations, annotators were not biased since they have

[11] https://www.copyright.gov/title17/92chap1.html#107.

the full access to all articles reporting on the same event but published by media of different ideology. Meanwhile, our developed systems were not designed to encode bias. In the training phase, we split the data on the story level, i.e., one story consisting of three articles from different ideologies, and we believe such training paradigm would help mitigate bias to a certain degree.

**Potential limitation.** Although balanced views are considered, the topic coverage in SEESAW is not exhaustive, and does not include other trending media or content of different modalities for expressing opinions, such as TV transcripts, images, and videos. Thus, the predictive performance of our developed system may still be under investigated.

In conclusion, there is no greater than minimal risk/harm introduced by either our dataset SEESAW or our developed novel system using it. However, to discourage misuse of SEESAW or stance detection related systems, we will always warn users that systems' outputs are for informational purpose only and users should always resort to the broader context to reduce the risk of absorbing biased information.

## References


Abeer Aldayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Inf. Process. Manag.*, 58:102597.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931. Association for Computational Linguistics.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885. Association for Computational Linguistics.

Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *ArXiv*, abs/1607.06450.

Matthew A Baum and Tim Groeling. 2008. New media and the polarization of american political discourse. *Political Communication*, 25(4):345–365.

Daniel Beck, Gholamreza Haffari, and Trevor Cohn. 2018. Graph-to-sequence learning using gated graph neural networks. In *ACL*.

Yochai Benkler, Robert Faris, and Hal Roberts. 2018. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

James Campbell. 2018. *Polarized: Making Sense of a Divided America*. Princeton University Press.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. *ArXiv*, abs/2010.00904.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557. Association for Computational Linguistics.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724. Association for Computational Linguistics.

Alyt Damstra, Mark Boukes, and Rens Vliegenthart. 2020. To Credit or to Blame? The Asymmetric Impact of Government Responsibility in Economic News. *International Journal of Public Opinion Research*, 33(1):1–17.

Wouter De Nooy and Jan Kleinnijenhuis. 2013. Polarization in the media during an election campaign: A dynamic network model predicting support and attack among political actors. *Political Communication*, 30(1):117–138.

Lingjia Deng and Janyce Wiebe. 2015. MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328. Association for Computational Linguistics.

Daniel Diermeier, Jean-François Godbout, Bei Yu, and Stefan Kaufmann. 2012. Language and ideology in congress. *British Journal of Political Science*, 42(1):31–55.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention. In *IJCAI*.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168. Association for Computational Linguistics.

Matthew Gentzkow and Jesse M. Shapiro. 2006. Media bias and reputation. *The Journal of political economy*, 114(2):280–316.

Matthew Gentzkow, Jesse M. Shapiro, and Daniel F. Stone. 2015. Chapter 14 - media bias in the marketplace: Theory. In Simon P. Anderson, Joel Waldfogel, and David Strömberg, editors, *Handbook of Media Economics*, volume 1 of *Handbook of Media Economics*, pages 623–645. North-Holland.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611. Association for Computational Linguistics.

Matt Grossmann and David A Hopkins. 2016. *Asymmetric politics: Ideological Republicans and group interest Democrats*. Oxford University Press.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503. Association for Computational Linguistics.

Carlee Beth Hawkins and Brian A Nosek. 2012. Motivated independence? implicit party identity predicts political judgments among self-proclaimed independents. *Personality and Social Psychology Bulletin*, 38(11):1437–1452.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

James Hong, Will Crichton, Haotian Zhang, Daniel Y. Fu, Jacob Ritchie, Jeremy Barenholtz, Ben Hannel, Xinwei Yao, Michaela Murray, Geraldine Moriba, Maneesh Agrawala, and Kayvon Fatahalian. 2021. Analysis of faces in a decade of us cable tv news. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5094–5107, Online. Association for Computational Linguistics.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *ICLR*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Ezra Klein. 2020. *Why We're Polarized*. Simon and Schuster.

Matthew Levendusky. 2013. Partisan media exposure and attitudes toward the opposition. *Political communication*, 30(4):565–581.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021a. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365. Association for Computational Linguistics.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2021b. Improving stance detection with multi-dataset learning and knowledge distillation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6332–6345, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weizhi Liao, Yu Wang, Yanchao Yin, Xiaobing Zhang, and Pan Ma. 2020. Improved sequence generation model for multi-label classification via cnn and initialized fully connection. *Neurocomputing*, 382:188–195.

Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 2021. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3152–3157. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. POLITICS: pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374.

Monty G. Marshall. 2005. Political conflict, measurement of. In Kimberly Kempf-Leonard, editor, *Encyclopedia of Social Measurement*, pages 89–98. Elsevier, New York.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952. European Language Resources Association (ELRA).

David Moss. 2017. *Democracy: A Case Study*. Belknap Press of Harvard University Press, 2017.

Kunwoo Park, Zhufeng Pan, and Jungseock Joo. 2021. Who blames or endorses whom? entity-to-entity directed sentiment extraction in news text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4091–4102. Association for Computational Linguistics.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge stage 1 (fnc-i): Stance detection.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27:443–460.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *NAACL*.

Dan Su, Yan Xu, Wenliang Dai, Ziwei Ji, Tiezheng Yu, and Pascale Fung. 2020. Multi-hop question generation with graph convolutional network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4636–4647, Online. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2020. X -stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference and the 16th Conference on Natural Language Processing, SwissText/KONVENS 2020, Zurich, Switzerland, June 23-25, 2020 [online only]*, volume 2624 of *CEUR Workshop Proceedings*.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Michael Welch, Melissa Fenwick, and Meredith Roberts. 1998. State managers, intellectuals, and the media: A content analysis of ideology in experts' quotes in feature newspaper articles on crime. *Justice quarterly*, 15(2):219–241.

Robert West, Hristo S Paskov, Jure Leskovec, and Christopher Potts. 2014. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics*, 2:297–310.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 23–30. Association for Computational Linguistics.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429. Association for Computational Linguistics.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Xiang Yue, Xinliang Frederick Zhang, Ziyu Yao, Simon M. Lin, and Huan Sun. 2021. Cliniqg4qa: Generating diverse questions for domain adaptation of clinical question answering. *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 580–587.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based

sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510. Association for Computational Linguistics.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models for question answering. *CoRR*, abs/2201.08860.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3901–3910. Association for Computational Linguistics.

Yiwei Zhou, Alexandra Ioana Cristea, and Lei Shi. 2017. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *WISE*.

## A Annotation Guideline

**Step 1: Entity annotation:** First, read the entire article and list the entities as well as their corresponding types. *Main entities* are the major subjects, objects, or participants involved in the main events described in the article. *Salient entities* broadly refer to political or notable figures that appear in the news stories even if they are not the main ones, including public figures, celebrities, or other important people, events, or subjects. Entities are always nominals (i.e., nouns or noun phrases), with examples and corresponding types listed below.

**Entity Name** must be a span of a text (please copy-and-paste from the text and stay with the surface form). If multiple versions exist, please use the most formal and complete span in the article. For example, the entity name for "Mayor Ben Zahn II", "Ben Zahn", "Zahn" or "he" (when referring to the entity) should be "Mayor Ben Zahn III".

**Entity Types** are from the set of {People, Organization, Place, Event, Religion, Topic, Other}.

**Step 2: Sentiment annotation:** Next, read one sentence at a time and for each sentence annotate the sentiment held by one mentioned entity (subject) towards another entity (object) in the triplet format of <subject, sentiment, object>. Either the subject, the object, or both must be in the entity list annotated in Step 1. Do not annotate relationships where neither subject nor object is in the entity list, but feel free to add to these list if you discover any entities you missed in step 1. If there are multiple triplets in a sentence, please annotate all fairly clearly sentiments in that sentence. If a sentence does not contain any triplet with at least one entity, leave it blank.

**Subject or Object values** are from {entities, "Not in the list", and "None"}

**"Not in the list"** should be used when a subject or object does not appear in the Entity list from Step 1. Note that when neither subject nor object appear in the list, the triplet should not be coded.

**"None"** is used in cases where the subject or the object does not exist, or can not be identified.

**Sentiment values** are from the {Positive, Negative}.

Note: Subject and object must be 1) explicitly mentioned in the text or 2) implicitly referred to by a pronoun (e.g., he, his) or by their titles (e.g., US President => Joe Biden), or 3) can be straightforwardly inferred from the context, e.g., the speaker identity is mentioned in previous sentences.

| News topic | # of news stories |
|---|---|
| Elections | 30 |
| Politics | 16 |
| White House | 15 |
| US House | 11 |
| US Senate | 10 |
| Immigration | 10 |
| Violence in America | 7 |
| Federal Budget | 7 |
| Gun Control and Gun Rights | 7 |
| Healthcare | 6 |
| US Congress | 5 |
| Coronavirus | 5 |
| Supreme Court | 4 |
| Justice Department | 4 |
| National Security | 4 |
| National Defense | 4 |
| State Department | 3 |
| Economic Policy | 3 |
| Terrorism | 3 |
| Economy and Jobs | 3 |
| LGBT Rights | 3 |
| Labor | 2 |
| Holidays | 2 |
| Race and Racism | 2 |
| Nuclear Weapons | 2 |
| FBI | 2 |
| Justice | 2 |
| Sexual Misconduct | 2 |
| Abortion | 2 |
| Education | 2 |
| Impeachment | 2 |
| Free Speech | 2 |
| Treasury | 2 |
| Republican Party | 1 |
| Religion and Faith | 1 |
| Fake news | 1 |
| Campaign Finance | 1 |
| Inequality | 1 |
| Donald Trump | 1 |
| Homeland Security | 1 |
| US Military | 1 |
| Public Health | 1 |
| Criminal Justice | 1 |
| Voting Rights and Voter Fraud | 1 |
| Joe Biden | 1 |
| NSA | 1 |
| Veterans Affairs | 1 |
| Cybersecurity | 1 |
| World | 1 |
| Middle East | 1 |
| Family and Marriage | 1 |
| Taxes | 1 |
| Total | 203 |

Table A1: Number of news stories for each topic in SEESAW.

| Media outlet | # of articles | Media-level ideology |
|---|---|---|
| Washington Times | 100 | Far Right |
| CNN (Online News) | 58 | Far Left |
| New York Times (News) | 52 | Lean Left |
| HuffPost | 51 | Far Left |
| Washington Post | 45 | Lean Left |
| Politico | 44 | Lean Left |
| USA TODAY | 38 | Lean Left |
| NPR (Online News) | 32 | Center |
| Newsmax (News) | 32 | Far Right |
| Townhall | 23 | Lean Right |
| The Hill | 23 | Central |
| Reuters | 21 | Central |
| BBC News | 15 | Central |
| Fox News (Online News) | 13 | Lean Right |
| Breitbart News | 11 | Lean Right |
| National Review | 11 | Lean Right |
| Vox | 10 | Far Left |
| The Guardian | 10 | Lean Left |
| Reason | 7 | Far Right |
| Christian Science Monitor | 7 | Center |
| Washington Examiner | 2 | Far Right |
| TheBlaze.com | 2 | Center |
| Wall Street Journal (News) | 1 | Center |
| Salon | 1 | Far Left |
| Total | 609 | - |

Table A2: Number of articles collected from each source and corresponding media-level ideology based on AllSides label.

**Step 3: Article-level entity-targeted sentiment annotation:** Next, having read the whole article, please annotate the overall article-level sentiments towards all listed entities based on your reading. If you are unsure about the sentiment, please mark it "Unknown".

**Article-level sentiment values** are {Very Positive, Positive, Slightly Positive, Neutral, Slightly Negative, Negative, Very Negative, Unknown}.

**Step 4: Entity ideology annotation:** Next, annotate the ideology of entities based on your reading. Entity ideologies must be determined or inferred based on a combination of your knowledge of the article, your knowledge of the overall political context, and your sentiment annotation. If there is no clear identifiable ideology associated with an entity, please mark it "Not Applicable".

**Entity ideology values** are from {Very liberal, Liberal, Slightly liberal, Moderate, Slightly conservative, Conservative, Very conservative, Not Applicable}.

**Step 5: Media-source ideology annotation:** Finally, attempt to estimate the ideology of the media organization that published this article. If you are unsure about the ideology, please mark it "Unknown".

**Media-source ideology values** are from {Very liberal, Liberal, Slightly liberal, Moderate, Slightly conservative, Conservative, Very conservative, Unknown}.

**Post-hoc conversion:** We further convert fine-grained labels obtained in steps 3 through step 5 to coarse-grained labels according to the nature of each task. For sentiment annotation, we convert them as 3-way labels. Specifically, we convert very positive and positive into one positive category, and similarly for very negative and negative. Then we merge slightly positive, neutral, and slightly negative into neutral. For ideological labels obtained in steps 4 and 5, in light of the 5-way annotation provided by AllSides, we also convert ours as 5-way labels by merging very liberal and liberal into liberal, and similarly for very conservative and conservative.

## B Details of Our Model

This section is supplementary to §4.3 and §4.4 in the main content, with more details about mathematical formulations and implementation details. Our framework takes as input a multi-sentence document, $\mathbf{x} = \{x_1, \ldots, x_{k+1}, \ldots, x_{k+t}, \ldots, x_L\}$, where the target sentence is in $\mathbf{x}$, i.e., $\tilde{\mathbf{x}} = \{x_{k+1}, \ldots, x_{k+t}\}$. Our model first generates a semantic graph $G$ as described in §4.1. $\mathbf{x}$ and $G$ are consumed by BART encoder (Lewis et al., 2020) and graph encoder (§4.2) separately, producing token representation, $\mathbf{H_T} \in \mathbb{R}^{m \times L}$, and node representations, $\mathbf{H_G} \in \mathbf{R}^{m \times N}$, where $N$ denotes the number of nodes in graph $G$. Finally, stance triplets are generated by our decoder using improved *in-parallel attention and information fusion mechanisms* (§4.3). Moreover, we inject Wikipedia knowledge to support the identification of relations between entities, especially to provide additional information which is not present in texts, e.g., party affiliations and geopolitical relations (§4.5).

### B.1 Decoder

We decode with our improved multi-source fused decoder, improved upon Transformer decoder (Vaswani et al., 2017), to enable reasoning over both information sources: text and graph.

The key difference between the vanilla Transformer decoder and ours is the *in-parallel cross-attention layer* which allows better integration of knowledge encoded in the two heterogeneous sources. Concretely, the cross attention to the text is formulated as:

$$\mathbf{z}_T = \text{LayerNorm}(\mathbf{z} + \text{Attn}(\mathbf{z}, \mathbf{H}_T)) \quad (2)$$

| | Full | | | Aspect | | |
|---|---|---|---|---|---|---|
| | Acc. | F1 | Acc.Any | src-s | s-tgt | src-tgt |
| **Baselines (No Graph)** | | | | | | |
| Sentence (Sen) | $7.26_{\pm 0.07}$ | $10.35_{\pm 0.31}$ | $12.39_{\pm 0.35}$ | $36.66_{\pm 0.50}$ | $23.81_{\pm 0.40}$ | $14.88_{\pm 0.37}$ |
| Sen + Context (Ctx) | $9.66_{\pm 0.18}$ | $14.08_{\pm 0.20}$ | $16.87_{\pm 0.24}$ | $45.24_{\pm 0.82}$ | $27.79_{\pm 0.35}$ | $20.01_{\pm 0.42}$ |
| Sen + Ctx + Entities | $11.32_{\pm 0.26}$ | $16.00_{\pm 0.34}$ | $19.15_{\pm 0.41}$ | $47.43_{\pm 1.16}$ | $30.03_{\pm 0.45}$ | $23.18_{\pm 0.34}$ |
| **Pipeline Models (Ours)** | | | | | | |
| Graph | $12.03_{\pm 0.58}$ | $15.77_{\pm 0.84}$ | $19.86_{\pm 0.83}$ | $46.60_{\pm 0.95}$ | $31.07_{\pm 0.65}$ | $23.19_{\pm 0.79}$ |
| + Oracle Entities | $31.84_{\pm 0.58}$ | $35.58_{\pm 0.76}$ | $44.87_{\pm 0.69}$ | $66.33_{\pm 1.08}$ | $55.50_{\pm 0.42}$ | $53.82_{\pm 0.61}$ |
| **End-to-end Models (Ours)** | | | | | | |
| Graph (seq. attn.; Cao and Wang, 2021) | $12.97_{\pm 0.34}$ | $17.22_{\pm 0.38}$ | $20.62_{\pm 0.46}$ | $50.58_{\pm 1.19}$ | $32.45_{\pm 0.66}$ | $24.76_{\pm 0.60}$ |
| Graph | $13.62_{\pm 0.23}$ | $18.12_{\pm 0.30}$ | $21.78_{\pm 0.37}$ | $51.01_{\pm 0.80}$ | $32.65_{\pm 0.41}$ | $\mathbf{26.08}_{\pm 0.43}$ |
| + Multitask | $13.34_{\pm 0.22}$ | $18.16_{\pm 0.69}$ | $21.77_{\pm 0.84}$ | $\mathbf{52.10}_{\pm 0.88}$ | $32.06_{\pm 0.84}$ | $26.07_{\pm 0.75}$ |
| + Wiki | $\mathbf{13.74}_{\pm 0.21}$ | $\mathbf{18.24}_{\pm 0.26}$ | $\mathbf{21.87}_{\pm 0.31}$ | $51.41_{\pm 0.55}$ | $\mathbf{32.69}_{\pm 0.69}$ | $25.94_{\pm 0.46}$ |

Table A3: Results on SEESAW for E2E stance detection task, and breakdown of accuracy scores by aspects (average of 5 runs). Best results without oracle entities are in **bold**. Our graph-augmented model with Wikipedia knowledge performs the best on 4 out of 6 metrics, indicating the effectiveness of encoding knowledge.

| | Accuracy | Macro F1 |
|---|---|---|
| BART (Lewis et al., 2020) | $86.32_{\pm 0.71}$ | $77.53_{\pm 0.71}$ |
| POLITICS (Liu et al., 2022) | $86.33_{\pm 0.83}$ | $77.48_{\pm 1.24}$ |
| DSE2QA (Park et al., 2021) | $87.78_{\pm 0.56}$ | $\mathbf{79.90}_{\pm 0.80}$ |
| Our model | $\mathbf{87.79}_{\pm 0.37}$ | $79.01_{\pm 0.66}$ |

Table A4: Results on stance-only prediction for specified pairwise entities (average of 5 runs). Our model performs on par with state-of-the-art models in stance detection tasks (POLITICS and DSE2QA). Our model performs on par with SOTA discriminative models.

where $\mathbf{z}$ denotes the output from the self-attention layer, $\mathbf{H_T}$ is the token representations out of the text encoder, and $Attn$ denotes the cross-attention mechanism as in Vaswani et al. (2017). We can compute $\mathbf{z_G}$ in a similar manner that attends node representations ($\mathbf{H_T}$) from the graph encoder.

$$\mathbf{z}_G = \text{LayerNorm}(\mathbf{z} + \text{Attn}(\mathbf{z}, \mathbf{H}_G)) \quad (3)$$

where $\mathbf{z}$ denotes the output from the same self-attention layer, $\mathbf{H_G}$ is the node representations out of the graph encoder.

Our *information fusion* module enables the information interaction between textual ($\mathbf{z_T}$) and graph ($\mathbf{z_G}$) hidden states, to obtain the fused representation, $\mathbf{z}'$. We implement the following two operations for information fusion: (1) addition, i.e., $\mathbf{z}' = \mathbf{z}_T + \mathbf{z}_G$, and (2) gating mechanism between $\mathbf{z}_T$ and $\mathbf{z}_G$ similar to (Zhao et al., 2018), as formulated below:

$$\mathbf{z_f} = \text{GELU}(\mathbf{W^f}[\mathbf{z_T}; \mathbf{z_G}] + \mathbf{b_f}) \quad (4)$$

$$\lambda = \text{sigmoid}(\mathbf{W}^\lambda[\mathbf{z_T}; \mathbf{z_G}] + \mathbf{b}_\lambda) \quad (5)$$

$$\mathbf{z} = \boldsymbol{\lambda} \odot \mathbf{z_f} + (1 - \boldsymbol{\lambda}) \odot \mathbf{z_T} \quad (6)$$

where $\odot$ is element-wise product, and $\mathbf{W}^*$ and $\mathbf{b}_*$ are learnable. $\boldsymbol{\lambda}$ here denotes the learnable gate vector. The selection of operation is decided by the downstream task. Specifically, in experiments we use addition for task A and gating mechanism for task B.

### B.2 Training Objectives

We adopt the cross entropy (CE) training objective that minimizes the following loss for model training.

$$\mathcal{L}_{stance} = -\sum_{(\mathbf{x},\mathbf{y})\in D} \log p(\mathbf{y}|\mathbf{x}) \quad (7)$$

where the reference $\mathbf{y}$ is a sequence of ground-truth stance triplet(s), sorted by their entities' first occurrences in the target sentence $\tilde{\mathbf{x}}$, and $D$ denotes the training set. $\mathbf{x}$ is formatted as *"<s> [preceding context] <s> [target text] </s> [succeeding context] </s>"*, where *[·]* indicates placeholders. Optionally, extracted entities can be paired with document $\mathbf{x}$ and then fed into the text encoder, in the format of "$\mathbf{x}$ <s> <ENT> $E_1$ <ENT> $E_2$ ...". $\mathbf{y}$ is formatted as *"<ENT> [source] <ENT> [target] <STANCE> [stance]"*, where <·> is a separator and *[·]* is a placeholder.

**Variant with Node Prediction.** In addition to modeling entity interactions in the graph, we enhance the model by adding an auxiliary objective to predict the node salience, i.e., whether the corresponding entity should appear in the stance triplets $\mathbf{y}$ to

be generated. This is motivated by the observation that $G$ usually contains excessive entity nodes, only a small number of which are involved in sentiment expression in the target sentence. Specifically, for each *entity node* $E_i$, we predict its salience, i.e., $\hat{s}_i$, by applying affine transformation over its representation $\mathbf{h}_G$, followed by a sigmoid function

$$\hat{\mathbf{s}} = \text{sigmoid}(\mathbf{u}\mathbf{H}_\mathbf{G}^\mathbf{E}) \tag{8}$$

where $\hat{\mathbf{s}} = \{\hat{s}_1, \ldots, \hat{s}_N\}$ is the collection of all entity nodes, $\mathbf{H}_\mathbf{G}^\mathbf{E}$ is a matrix of node representations for entity nodes out of the graph encoder, and $\mathbf{u}$ is learnable during training.

We adopt the weighted binary cross entropy (BCE) training objective to minimize the loss, $\mathcal{L}_{node}$, over all *entity nodes*.

$$\mathcal{L}_{node} = -\sum_{s_i} w * s_i \log(\hat{s}_i) + (1 - s_i)\log(1 - \hat{s}_i) \tag{9}$$

where $w$ controls loss weights on positive samples, and $s_i$ denotes the occurrence of entity node $E_i$ in the ground-truth stance triplet $\mathbf{y}$.

Finally, when the node prediction module is enabled, the overall loss for the **multitask** learning setup is $\mathcal{L}_{multi} = \mathcal{L}_{stance} + \mathcal{L}_{node}$.

## C SEESAW Annotation Quality Control

We hold meetings on a weekly basis and give annotators timely feedbacks to resolve annotation disagreements and iteratively improve annotation quality. We randomly sample $10\%$ news stories and have them annotated by multiple people. We first evaluate on the overlapping ratio between a pair of entity sets extracted by two different people. The overlapping ratio is $55.5\%$ after cross-document entity resolution is conducted. Though the overlapping ratio is not high, we do find that entities captured by one but not both can help complement one another's annotation. Next, for sentiment annotation, we compute the agreement level by comparing two annotators' sentiment annotations on items that are annotated by both. We reach $97\%$ agreement level, showing high quality of our SEESAW. Further, a simple unadjusted agreement between AllSides media-level ideology label and annotator's perception of the article's ideological leaning is 0.77 out of 1.0.[12]

[12]We consider the difference within one level as correct matching, e.g., *Far Left (0)* and *Lean Left (1)* are matched.

... In her seven-page opinion, Justice Sotomayor wrote that the Trump administration had become too quick to run to the Supreme Court after interim losses in the lower courts. "Claiming one emergency after another, the government has recently sought stays in an unprecedented number of cases, demanding immediate attention and consuming limited court resources in each," she wrote. "And with each successive application, of course, its cries of urgency ring increasingly hollow." ...

`[0] Sonia Sotomayor NEG Donald Trump`

**Sent.:** <Someone> NEG Affordable Care Act

**Sent. + Cxt.:** Supreme Court of the United States NEG Donald Trump

**Sent. + Cxt. + Ent.:** Donald Trump NEG <Someone>

**Graph model** (ours): Sonia Sotomayor NEG Donald Trump

**Graph model + Wiki** (ours): Sonia Sotomayor NEG Donald Trump

... The actions drew charges of racism because more than 200,000 Black Michiganders would have their votes disallowed by **such an action**. Palmer's comment that she would be willing to certify results in Detroit's suburbs - which experienced some of the same clerical errors that Detroit did - but not in Detroit, was seen as particularly outrageous. Both Palmer and Hartmann changed their votes on certification to "yes" Tuesday after strong criticism and heartfelt appeals from citizens participating in the board meeting over Zoom. But in the **affidavits** signed late Wednesday, both said they feel they made those votes under pressure and false pretenses and would now like to rescind their votes. ...

`[0] <Someone> NEG affidavit`

**Sent.:** <Author> NEG Barack Obama

**Sent. + Cxt.:** <Someone> NEG Wayne County Board of Canvassers

**Sent. + Cxt. + Ent.:** <Someone> NEG Wayne County, Michigan

**Graph model** (ours): <Someone> NEG Wayne County, Michigan

**Graph model + Wiki** (ours): <Someone> NEG Wayne County, Michigan

Figure A1: Additional error analysis on two more test samples. The underlined sentence is the target sentence. In the top example, both of our model are able to capture the correct stance triplet on this challenging sample while the baselines all fail, showing the power of graph modeling of entity interaction. In the bottom example, none of the five models get it right. More powerful models should be developed to encode broader context and knowledge. For example, event-level co-reference resolution seems to be imperative in order to understand the connection between *affidavit* and the phrase "such an action".

## D Reproducibility Checklist

For all experiments, we use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 1e-5 and fine-tune up to 15 epochs. The batch size of all baselines and our models are 4. The gradient is clipped when its norm exceeds 5. We select the best model for each method using the accumulated loss on the dev set. In decoding, the batch size is 1. We also enable learning rate decay with a patience of 200 steps. The early stop is also enabled with a patience of $1,600$ steps. For all these other hyperparameters, we keep the default values.